Early detection of cervical cancer

Snoek, B.C.

2019

document version
Publisher's PDF, also known as Version of record

CHAPTER 6

IDENTIFICATION AND VALIDATION OF A 3-GENE METHYLATION CLASSIFIER FOR HPV-BASED CERVICAL SCREENING ON SELF-SAMPLES

Wina Verlaat
Barbara C. Snoek
Daniëlle A.M. Heideman
Saskia M. Wilting
Peter J.F. Snijders
Putri W. Novianti
Annina P. van Splunter
Carel F.W. Peeters
Nienke E. van Trommel
Leon F.A.G. Massuger
Ruud L.M. Bekkers
Willem J.G. Melchers
Folkert J. van Kemenade
Johannes Berkhof
Mark A. van de Wiel
Chris J.L.M. Meijer
Renske D.M. Steenbergen

Clinical Cancer Research 2018; 24(14): 3456-3464

ABSTRACT

Background

Offering self-sampling of cervico-vaginal material for high-risk human papillomavirus (hrHPV) testing is an effective method to increase the coverage in cervical screening programmes. Molecular triage directly on hrHPV-positive self-samples for colposcopy referral opens the way to full molecular cervical screening. Here, we set out to identify a DNA methylation classifier for detection of cervical precancer (CIN3) and cancer, applicable to lavage and brush self-samples.

Methods

We determined genome-wide DNA methylation profiles of 72 hrHPV-positive self-samples, using the Infinium Methylation 450K Array. The selected DNA methylation markers were evaluated by multiplex quantitative methylation-specific PCR (qMSP) in both hrHPV-positive lavage (n=245) and brush (n=246) self-samples from screening cohorts. Subsequently, logistic regression analysis was performed to build a DNA methylation classifier for CIN3 detection applicable to self-samples of both devices. For validation, an independent set of hrHPV-positive lavage (n=199) and brush (n=287) self-samples was analysed.

Results

Genome-wide DNA methylation profiling revealed 12 DNA methylation markers for CIN3 detection. Multiplex qMSP analysis of these markers in large series of lavage and brush self-samples yielded a 3-gene methylation classifier (*ASCL1*, *LHX8* and *ST6GALNAC5*). This classifier showed a very good clinical performance for CIN3 detection in both lavage (AUC=0.88; sensitivity=74%; specificity=79%) and brush (AUC=0.90; sensitivity=88%; specificity=81%) self-samples in the validation set. Importantly, all self-samples from women with cervical cancer scored DNA methylation-positive.

Conclusion

By genome-wide DNA methylation profiling on self-samples, we identified a highly effective 3-gene methylation classifier for direct triage on hrHPV-positive self-samples, which is superior to currently available methods.

INTRODUCTION

Organised cytology-based cervical screening programmes using physician-collected cervical scrapes have led to a substantial decrease in cervical cancer incidence and mortality in high-income countries[1]. However, a considerable subset of women does not attend cervical screening (non-attendees), which compromises the effectiveness of the screening programme[2]. Previous studies have shown that offering self-sampling of cervico-vaginal specimens (self-samples) for high-risk human papillomavirus (hrHPV) testing (hrHPV self-sampling) to non-attendees increases the attendance to cervical screening. Up to 30% of the invited non-attendees returned their self-sample to the laboratory for hrHPV testing[3–6]. Importantly, the diagnostic accuracy of hrHPV testing on self-samples for cervical intraepithelial neoplasia grade 3 and cervical cancer (CIN3+) is similar to hrHPV screening of physician-collected cervical scrapes[7,8]. Therefore, offering hrHPV self-sampling as an alternative to conventional scrapes has just been implemented in the new HPV-based cervical screening programme in the Netherlands. Partial substitution of hrHPV testing on physician-collected scrapes in cervical screening programmes by hrHPV self-sampling can be envisioned in the near future.

Although hrHPV testing has a higher sensitivity for CIN3+ compared with cytology, its 3% to 5% lower specificity for CIN3+ necessitates the use of a triage test to distinguish women with clinically relevant disease from those with irrelevant, transient hrHPV infections to prevent over-referral and overtreatment. Currently, cytology is the most widely accepted triage tool. Because cytology cannot be reliably performed on self-sampled material[9–11], women with hrHPV-positive self-samples need to visit a physician for an additional cervical scrape for cytology. This may lead to loss to follow-up, delay the diagnostic track and is less feasible in low-income countries given the lack of adequate infrastructure and limited number of trained practitioners[8,12,13]. Therefore, molecular triage testing directly applicable to self-sampled material from hrHPV-positive women is preferred.

We and others have shown that DNA methylation analysis of tumour-suppressor genes on self-samples is well feasible and effective to detect CIN3+ using quantitative methylation-specific PCR (qMSP)[12–16]. DNA methylation analysis has already shown competitive clinical performance versus other triage options in cervical scrapes, whereas improvements in performance on self-samples are conceivable. Previous findings have shown that DNA methylation markers originally discovered in tissue specimens and tested on hrHPV-positive cervical scrapes are not necessarily of clinical value when applied to hrHPV-positive self-samples[17]. This is likely due to the cellular composition of self-samples, which contain fewer disease-related cells. Therefore, self-samples may display distinct epigenetic signatures compared with physician-collected cervical specimens. Hence, DNA methylation marker discovery screens directly performed on self-samples are more likely to yield the most informative DNA methylation markers for hrHPV-positive self-samples.

In this study, we describe the identification and validation of a DNA methylation classifier for the detection of CIN3 and cervical cancer in hrHPV-positive self-samples. A genome-wide DNA methylation marker discovery for CIN3 detection was performed using the Infinium 450K BeadChip array to 72 hrHPV-positive self-samples from a screening cohort of non-attendees. The identified candidate DNA methylation markers were evaluated by multiplex qMSP in unique, large series of lavage-based (n=245; further referred to as "lavage self-samples") and brush-based (n=246; further referred to as "brush self-samples") self-samples from screening cohorts of non-attendees to build an optimal DNA methylation classifier for detection of CIN3 that is applicable to self-samples of both devices. The clinical performance of the obtained DNA methylation classifier was subsequently validated by multiplex qMSP on an independent series of lavage (n=199) and brush (n=287) self-samples.

MATERIALS AND METHODS

Clinical specimens

Discovery set: case-control series for DNA methylation marker discovery screen

For genome-wide DNA methylation marker discovery for CIN3 detection, hrHPV-positive lavage self-samples collected using the Delphi Screener (Delphi Bioscience) were obtained from a screening cohort of non-attendees (PROHTECT-1 trial[3]; NTR792; n=72; Fig. 1, Discovery screen). Detailed characteristics of study design, clinical specimens, inclusion criteria and follow-up procedures have been described previously[3]. Array data from a pilot experiment of 12 self-samples for power calculations revealed a ratio of 3 (hrHPV-positive controls) to 4 (CIN3) for proper marker discovery. Therefore, the discovery series comprised hrHPV-positive lavage self-samples from 29 control women, who either had histologic evidence of absence of CIN2+ (≤ CIN1) or displayed hrHPV clearance combined with normal cytology in follow-up (further referred to as hrHPV-positive controls; median age, 36; range, 31-56), and 39 cases histologically diagnosed with CIN3 (median age, 36; range, 31-62). Controls and cases were matched according to age and hrHPV type to the extent of sample availability. The hrHPV types in controls were eight HPV16, four HPV51, four HPV52, four HPV56, three HPV45, two HPV35, two HPV58, two HPV66, one HPV33 and one HPV39; the hrHPV types in CIN3 were 21 HPV16, six HPV31, four HPV52, three HPV33, three HPV56, two HPV51, two HPV68, one HPV18, one HPV35, one HPV39, one HPV45 and one HPV66. In addition, hrHPV-positive lavage self-samples from women histologically diagnosed with cervical squamous cell carcinoma (SCC; n=4) were included (median age, 49; range, 42-61). The hrHPV types in SCC were two HPV16, one HPV31 and one HPV45.

Building set: case-control series to build a DNA methylation classifier

To build a DNA methylation classifier for CIN3 detection, both hrHPV-positive lavage self-samples (n=245; PROHTECT-1 trial[3]; excluding samples used for the discovery screen) and brush self-samples collected using a Viba-Brush® (Rovers Medical Devices B.V., Oss, The Netherlands; n=246; PROHTECT-2 trial[4]; NTR1851) were obtained from screening cohorts of non-attendees who reached a study endpoint and all of which were not preselected (Fig. 1; building a DNA methylation classifier; Supplementary Fig. S1). Detailed characteristics of study design, clinical specimens, inclusion criteria and follow-up procedures have been described previously[4]. Available lavage self-samples of 214 hrHPV-positive controls (controls; median age, 41; range, 31-62) and 31 women histologically diagnosed with CIN3 (cases; median age, 36; range, 31-62) were included. Brush self-samples included 174 hrHPV-positive controls (controls; median age, 37; range, 30-62) and 72 women histologically diagnosed with CIN3 (cases; median age, 36; range, 31-61).

Validation set: independent series to validate the DNA methylation classifier

To validate the clinical performance of the DNA methylation classifier, independent series of both hrHPV-positive lavage (n=199) and brush (n=287) self-samples, all of which were not preselected, were used (Fig. 1; Validation of DNA methylation classifier; Supplementary Fig. S1). For lavage self-samples, hrHPV-positive samples collected using the Delphi Screener (Delphi Bioscience) were obtained from a screening cohort of non-attendees who reached a study endpoint in the PROHTECT-3 trial (methylation-arm; NTR2606)[12]. Detailed characteristics of study design, clinical specimens, inclusion criteria and follow-up procedures have been described previously[12]. Half of the available samples in this trial were randomly chosen for evaluation in the current study. These were supplemented with an independent series of four lavage self-samples from women with SCC who participated in the PROHTECT-1 trial[3].

The total lavage series comprised 134 hrHPV-positive controls (median age, 38; range, 33-63), 22 women with CIN2 (median age, 38; range, 33-58), 35 women with CIN3 (median age, 38; range, 33-48), seven women with SCC (median age, 48; range, 38-61) and one woman with adenocarcinoma (AdCA; age 33).

For brush self-samples, hrHPV-positive samples collected using the Evalyn® Brush (Rovers Medical Devices B.V., Oss, The Netherlands) were obtained from a screening cohort of non-attendees who reached a study endpoint in the PROHTECT-3B trial (NTR3350)[18]. Detailed characteristics of study design, clinical specimens, inclusion criteria and follow-up procedures have been described previously[18]. These were supplemented with an independent series of four brush self-samples from women with SCC and one brush self-sample from a woman with adenocarcinoma in situ (ACIS) who participated in the PROHTECT-2 trial[4] and seven brush self-samples from women with SCC and five brush self-samples from women with AdCA who visited the gynaecology clinic (METC15.1468/X15MET study). The total brush series comprised 178 hrHPV-positive controls (median age, 39; range, 33-63), 28 women with CIN2 (median age, 38; range, 33-53), 56 women with CIN3 (median age, 38; range, 33-59), 16 women with SCC (median age, 44; range, 29-75), one woman with ACIS (age 41) and eight women with AdCA (median age, 44; range, 27-62).

This study followed the ethical guidelines of the Institutional Review Board of VU University Medical Centre and Antoni van Leeuwenhoek Hospital/Netherlands Cancer Institute. All participants in the PROHTECT and X15MET trials gave informed consent.

Infinium HumanMethylation450 BeadChip and data preprocessing

Before application, quality of the DNA was assessed by Qubit BR dsDNA measurement and visual evaluation of DNA integrity on an agarose gel. Genome-wide DNA methylation profiling was performed by Infinium HumanMethylation450 BeadChip (Illumina). Data are available

from the NCBI Gene Expression Omnibus (GEO) through series accession number GSE99511 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE99511). Data were preprocessed and checked for sample and probe quality. Four samples (one hrHPV-positive control and three CIN3) were excluded based on the correlation heatmap results (Supplementary Fig. S2). For further data preprocessing, see Supplementary Methods.

hrHPV and DNA methylation testing

For sample processing, hrHPV testing and DNA methylation analysis, see Supplementary Methods and Supplementary Table S1. HrHPV positivity was determined for all samples. HrHPV genotypes were defined in a subset of the classifier building set only. In each multiplex qMSP assay, three targets and the housekeeping gene β-actin (*ACTB*) were combined as described before[19]. Target DNA methylation values were normalised to reference gene *ACTB* and the calibrator using the comparative C_t method ($2^{-\Delta\Delta Ct}$ x 100) to obtain $\Delta\Delta C_t$ ratios[20]. The $\Delta\Delta C_t$ ratios were square root-transformed. Only samples for which sufficient DNA material was available and which achieved an *ACTB* C_t value < 30 were included.

Statistical analyses

Discovery screen: genome-wide DNA methylation array data

After preprocessing of Infinium data, we applied adaptive group-regularised logistic ridge regression (GRridge)[21]. We incorporated auxiliary information (referred to as co-data) in building the GRridge classification model, namely *P* values from a similar study in cervical tissue specimens using the same array platform (Farkas and colleagues[22]) and standard deviation (sd) of each probe in the current dataset. Using informative co-data has been shown to enhance the identification of valuable markers in rather impure samples, such as self-samples (Supplementary Fig. S3). More details regarding the GRridge model by incorporating such information are provided in Supplementary Methods and elsewhere[21]. *Post hoc* forward selection was applied to the GRridge model to render a model of DNA methylation markers. The performance of the GRridge model was visualised by a receiver operating characteristics (ROC) curve, obtained by leave-one-out cross-validation, and quantified by area under the curve (AUC). Predicted probabilities, representing the risk for an underlying CIN3, were calculated using the GRridge model. Hierarchical clustering of the 28 DNA methylation markers was performed to further select the genes that were most discriminative between CIN3 and hrHPV-positive controls.

Building and validation of DNA methylation classifier: qMSP data

To compare DNA methylation levels between two groups (hrHPV-positive controls and CIN3), the Wilcoxon rank-sum test (two-sided) was applied on the square root-transformed $\Delta\Delta C_t$ ratios. Statistical significance was set at $P < 0.05$.

To build a DNA methylation classifier, classical logistic regression analysis was performed on qMSP data to select relevant DNA methylation markers for CIN3 detection in both lavage and brush self-samples (detailed description in Supplementary Methods and Supplementary Fig. S4). In brief, logistic regression analysis followed by stepwise selection and backward elimination was performed on the combination of lavage and brush self-sample datasets (to encourage overlap) to obtain an initial marker panel of two DNA methylation markers for both self-sample types. Forward selection on the separate lavage and brush datasets suggested the addition of a third DNA methylation marker, which was particularly relevant for the brush dataset, without harming the performance in the lavage dataset. Because DNA methylation in CpG islands has been shown to increase with age[23], we included age as a factor in the DNA methylation classifier. Supplementary Table S2 shows the *P* value and contribution (coefficient/sd) of age and the third DNA methylation marker *ST6GALNAC5* in the 3-gene methylation classifier. These two factors were included in the classifier because exclusion of age and *ST6GALNAC5* resulted in a lower performance in particularly the brush self-samples. Predicted probabilities and 95% confidence interval (CI) were calculated for all analysed samples using the logistic regression models of the DNA methylation classifier for lavage and brush self-samples. The clinical performance of the logistic regression models in both classifier building and validation sets was visualised by an ROC curve and evaluated by AUC calculation. The ROC curves show the sensitivity and specificity for the complete spectrum of different thresholds in predicted probabilities using the logistic regression models. A threshold was fixed for predicted probabilities corresponding to 80% specificity (lavage self-samples: 0.053; brush self-samples: 0.240) based on the classifier building set and subsequently evaluated in the independent validation set for CIN3 sensitivity and specificity. In addition, the DNA methylation classifier at a fixed threshold was applied on self-samples from women with CIN2, SCC and ACIS/AdCA to evaluate the positivity rates in these disease categories. A classification and regression tree (CART) algorithm, which renders a DNA methylation classifier using marker-based cutoffs, was built for comparison with the continuous values obtained by regression. For the details of the CART method, see Supplementary Methods.

RESULTS

An overview of the study design is given in Fig. 1.

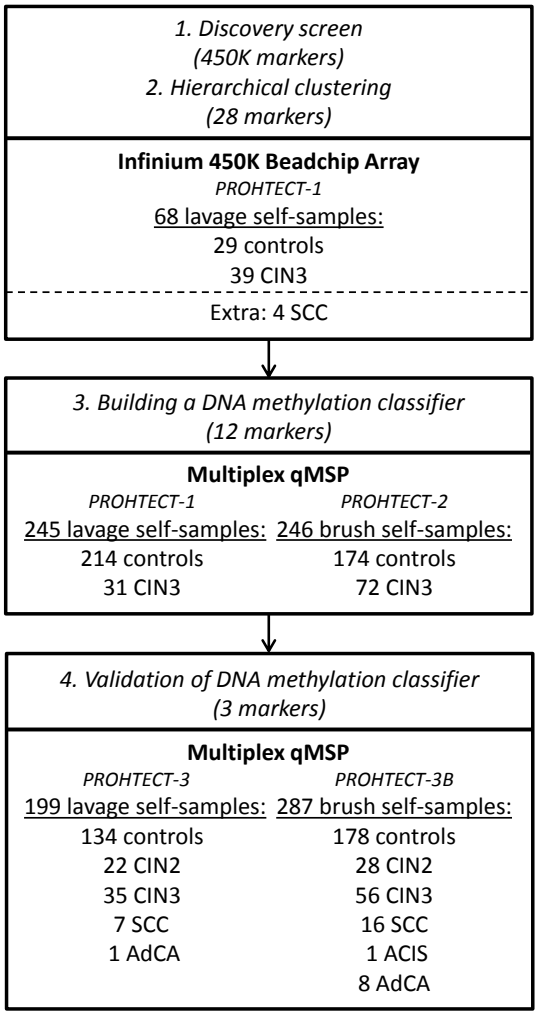


Figure 1. Experimental set-up of the study. All self-samples were obtained from screening cohorts of non-attendees, except seven SCC and five AdCA brush self-samples in the validation set.

Discovery of DNA methylation markers in hrHPV-positive self-samples

In total, we obtained 68 genome-wide DNA methylation profiles of hrHPV-positive lavage self-samples from a screening cohort of non-attendees, of which 64 (28 controls and 36 women with CIN3) were suitable to identify DNA methylation markers for CIN3 detection (Fig. 1; Discovery screen). Adaptive group-regularised ridge regression, GRridge[21] and variable selection on the DNA methylation profiles from women with and without CIN3 yielded a panel of 28 DNA methylation markers with discriminatory power for CIN3 (AUC of 0.77). Hierarchical clustering of all 28 DNA methylation markers showed that 12 methylated genes, that is, *ACAN*, *ASCL1*, *LHX8*, *MYADM*, *NRG3*, *RGS7*, *ST6GALNAC3*, *ST6GALNAC5*, *WDR17*, *ZNF582*, *ZNF583* and *ZNF781*, were mostly contributing to the discrimination of women with and without CIN3 (Fig. 1; Hierarchical clustering, Fig. 2A, Table 1; Supplementary Fig. S5). Evaluation of the DNA methylation profiling data from four hrHPV-positive lavage self-samples from women with SCC confirmed high DNA methylation levels for all these 12 DNA methylation markers (Fig. 2B; Supplementary Fig. S5).

Table 1. The 12 candidate DNA methylation markers from the discovery screen.

Infinium BeadChip probe	Chr.	Chr. location	Gene name
cg08272731	1	75602167	*LHX8*
cg14156405	1	241520286	*RGS7*
cg20707222	1	76540222	*ST6GALNAC3*
cg23243867	1	77334045	*ST6GALNAC5*
cg27486637	4	176987174	*WDR17*
cg10401879	10	83634276	*NRG3*
cg20718350	12	103352294	*ASCL1*
cg06675190	15	89346205	*ACAN*
cg13499300	19	54369556	*MYADM*
cg02763101	19	56904945	*ZNF582*
cg00796360	19	56915650	*ZNF583*
cg14587524	19	38183262	*ZNF781*

Chr.: Chromosome

Building a DNA methylation classifier using hrHPV-positive lavage and brush self-samples

Next, the 12 most discriminative DNA methylation markers from the discovery screen were further analysed using multiplex qMSP in large series of hrHPV-positive lavage self-samples (n=245) and brush self-samples (n=246) from women with and without CIN3 from two screening cohorts (Fig. 1; building a DNA methylation classifier). In both lavage and brush self-samples, all genes except *ACAN* (in lavage only; $P < 0.05$) showed significantly increased DNA methylation levels ($P < 0.001$) in self-samples from women with CIN3 compared with hrHPV-positive controls (Fig. 3).

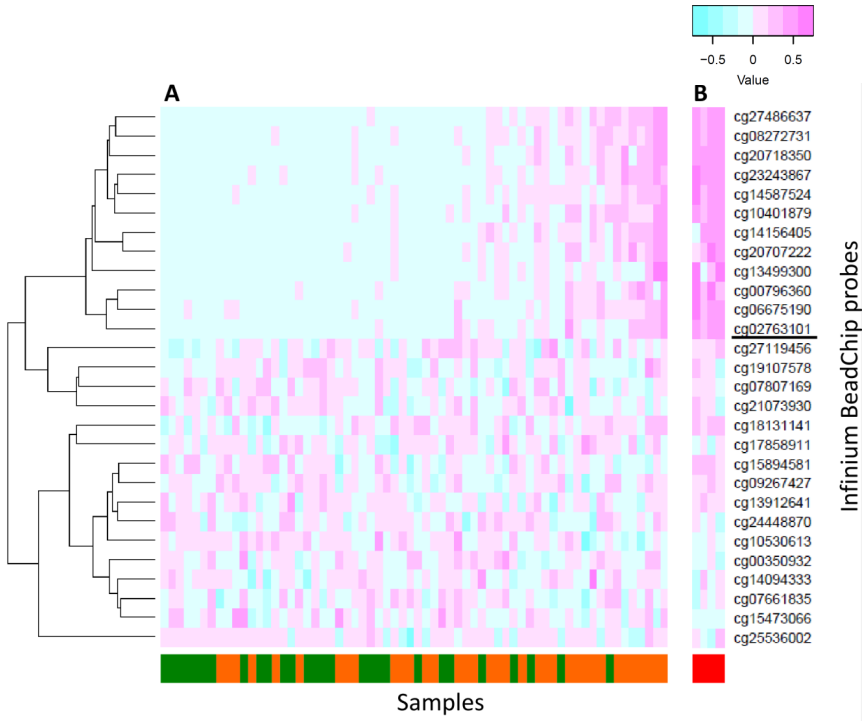


Figure 2. Heatmap of the 28 DNA methylation markers in the discovery screen. Hierarchical clustering of the 28 Infinium 450K BeadChip probes, each probe corresponds to a DNA methylation marker. Low (blue) to high (purple) DNA methylation levels (arcsine square root-transformed beta values) are displayed for each DNA methylation marker (cg numbers of the probes). (**A**) DNA methylation data of self-samples from hrHPV-positive controls (green; n=28) and from women with CIN3 (orange; n=36). The samples are ordered by predicted probability. The 12 DNA methylation markers above the black line showed the most discriminative DNA methylation profile between women with and without CIN3. (**B**) DNA methylation data of self-samples from women with SCC (red; n=4).

To build an optimal DNA methylation classifier for detection of CIN3, which is applicable to different self-sample types, logistic regression analysis followed by stepwise selection and backward elimination was performed on the combined dataset of lavage and brush self-sample qMSP results (see Materials and Methods, Supplementary Methods and Supplementary Fig. S4). This revealed a 3-gene methylation classifier for CIN3 detection in both self-sample types, consisting of *ASCL1*, *LHX8* and *ST6GALNAC5* (Supplementary Fig. S4; Supplementary Table S2). This 3-gene methylation classifier showed a very good clinical performance for CIN3 detection in both hrHPV-positive lavage (AUC=0.90) and brush (AUC=0.86) self-samples (Fig. 4A and B, black lines). At the threshold corresponding to a specificity of 80% in hrHPV-positive controls, 83% (25 of 30) of lavage self-samples and 76% (52 of 68) of brush self-samples from women with CIN3 were DNA methylation-positive (Supplementary Fig. S6).

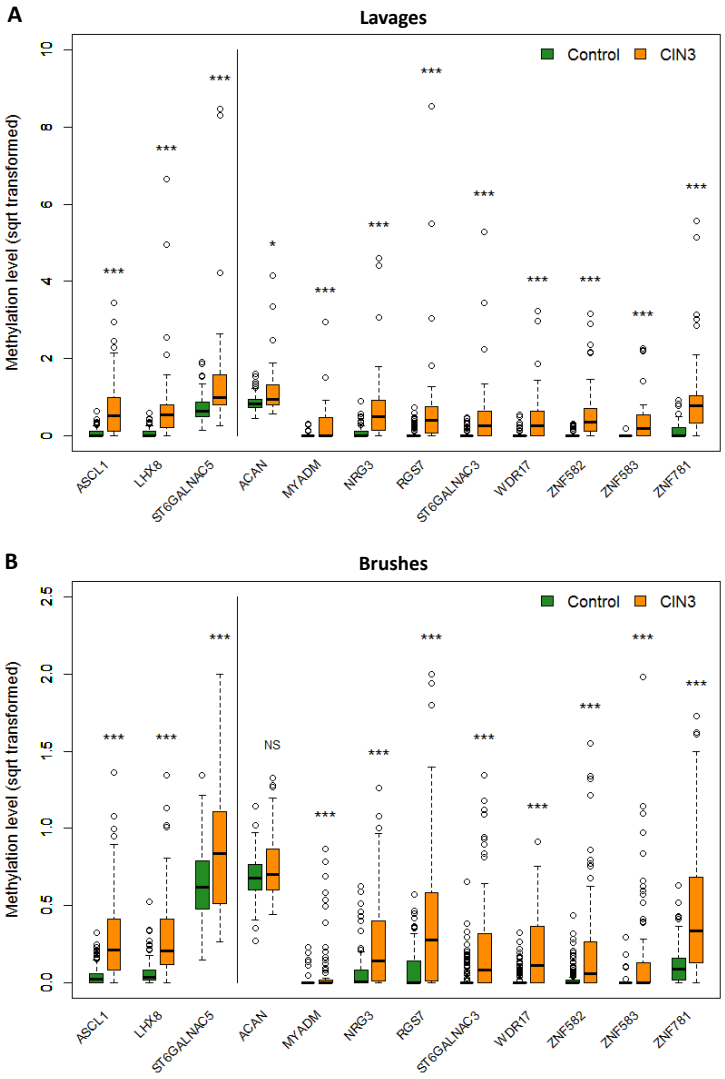


Figure 3. Differential DNA methylation levels of the 12 candidate methylation markers in hrHPV-positive self-samples. DNA methylation levels represented by the square root-transformed $\Delta\Delta C_t$ ratios (y axis) in (**A**) lavage self-samples from hrHPV-positive controls (n=214) and women with CIN3 (n=31; x axis) and (**B**) brush self-samples from hrHPV-positive controls (n=174) and women with CIN3 (n=72; x axis). The three genes left of the black line are included in the 3-gene methylation classifier. * $P < 0.05$; *** $P < 0.001$; NS: not significant.

Validation of DNA methylation classifier

To validate the clinical performance of the 3-gene methylation classifier, an independent, large series of hrHPV-positive lavage self-samples (n=199) and brush self-samples (n=287) was analysed by multiplex qMSP (Fig. 1; Validation of DNA methylation classifier). Solely hrHPV-positive controls and CIN3 from independent screening cohorts were used for validation of the 3-gene methylation classifier. This showed a comparable clinical performance for CIN3 detection as observed in the above-described classifier building set, in both hrHPV-positive lavage (AUC=0.88) and brush (AUC=0.90) self-samples (Fig. 4A and B, grey lines). The predefined threshold corresponding to an 80% specificity in the classifier building set (see above) was applied to this validation set. This resulted in a CIN3 sensitivity of 74% (26 of 35) in lavage self-samples and 88% (49 of 56) in brush self-samples, at 79% and 81% specificity in hrHPV-positive controls, respectively (Supplementary Fig. S6). To confirm these findings, we applied an alternative method (CART) on both lavage and brush self-samples, which rendered similar results to those shown here (Supplementary Methods; Supplementary Table S3; and Supplementary Figs. S7 and S8).

Furthermore, this validation series also comprised self-samples from women with CIN2 from a screening cohort. Fifty percent of these lavage self-samples (11 of 22) and brush self-samples (14 of 28) were DNA methylation-positive (Supplementary Fig. S6). Importantly, all 23 SCC (7 lavage self-samples and 16 brush self-samples; Supplementary Fig. S6), and all ACIS (1 brush self-sample) and AdCA (1 lavage self-sample and 8 brush self-samples) scored DNA methylation-positive (Supplementary Fig. S9).

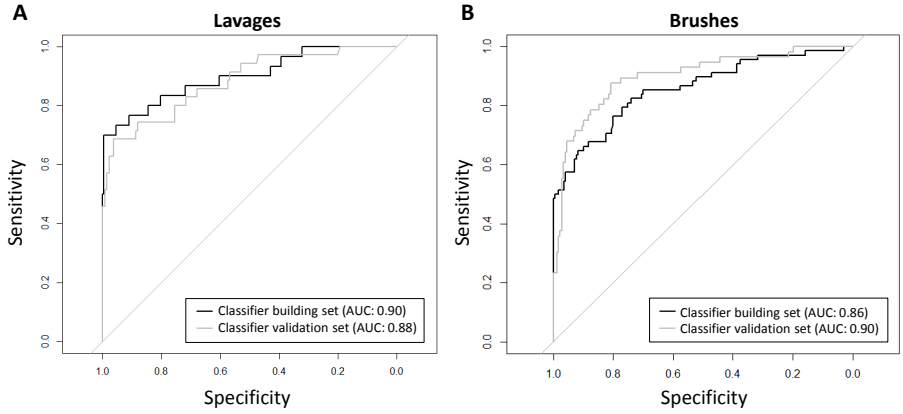


Figure 4. Clinical performance of the 3-gene methylation classifier for CIN3 detection in hrHPV-positive lavage and brush self-samples. ROC curve and AUC of the 3-gene methylation classifier for CIN3 detection in (**A**) lavage and (**B**) brush self-samples in the classifier building set (grey) and validation set (black).

DISCUSSION

Here, we identified a DNA methylation classifier consisting of three methylated gene promoters, *ASCL1*, *LHX8* and *ST6GALNAC5*, for the detection of CIN3 and cervical cancer in hrHPV-positive self-samples and validated the clinical performance in large series of both cervical lavage and brush self-samples from independent screening cohorts of non-attendees.

Previous publications showed that CIN lesions detected by DNA methylation analysis do not completely overlap with those detected by cytology[24]. In fact, DNA methylation analysis tends to preferably detect cervical cancer and advanced high-grade precursor lesions, defined as CIN2/3 associated with a persistent hrHPV infection of ≥ 5 years. Women with advanced CIN2/3 are presumed to have a high short-term progression risk to cancer and are therefore in need of immediate referral and treatment[24,25]. Cytology on the other hand detects both early and advanced CIN lesions with a moderate sensitivity of 65% to 80% and cannot be reliably applied to self-samples, requiring a visit to the physician[9–11]. DNA methylation markers are applicable on self-samples and have the potential to reduce the risk for undetected cervical cancers and advanced CIN2/3. On the contrary, women with a negative DNA methylation marker test would have a low short-term cancer progression risk, indicating that immediate colposcopy referral is unnecessary. To prevent over-referral and overtreatment in HPV-based self-sampling, direct triage testing by DNA methylation markers in self-sampled material enables the identification of only those hrHPV-positive women with clinically relevant disease who are in need of treatment and it allows for full molecular cervical self-screening.

This is the first study performing a discovery screen directly on self-samples, which allowed us to define the most optimal DNA methylation classifier for direct molecular triage testing on hrHPV-positive self-sampled material. Our 3-gene methylation classifier showed a very good and reproducible clinical performance for detection of CIN3 in both hrHPV-positive lavage (classifier building set AUC=0.90; classifier validation set AUC=0.88) and brush (classifier building set AUC=0.86; classifier validation set AUC=0.90) self-samples. This indicates that it represents a universal triage test for both self-sample devices. Furthermore, the combined analysis of the 3-gene methylation classifier and a reference gene in a single multiplex assay saves material, costs and time and allows for (semi)high-throughput screening.

To select the most discriminatory DNA methylation markers for CIN3 from our discovery screen on hrHPV-positive self-samples, which are rather impure due to an overrepresentation of non-disease-related cells, we applied our recently proposed GRridge model[21]. This method enables objective use of co-data and was shown to potentially outperform other prediction methods (Supplementary Fig. S3)[26]. In particular publicly available DNA methylation data from relatively pure cervical tissue specimens, obtained by the same array platform, proved to be

useful co-data[22]. The validity of this approach is supported by the identification of the three DNA methylation classifier genes that have all been previously described in DNA methylation studies on cervical cancer[22,27,28]. The combination of GRridge (on array data) and classical logistic regression analysis (on qMSP data) enabled us to build a highly discriminative methylation classifier for CIN3 detection consisting of *ASCL1*, *LHX8* and *ST6GALNAC5*. The narrow range of the 95% CI of the predicted probabilities (i.e. the methylation classifier value; range, 0-1) in both lavage and brush self-samples supports a good representation of the disease state (case vs. control) in the population by the 3-gene methylation classifier (Supplementary Fig. S10). Comparison of the three markers in HPV16-positive self-samples to self-samples positive for other hrHPV types (non-HPV16), in the subset of samples with HPV typing information, revealed no significant difference in DNA methylation levels in both lavage and brush self-samples, except for *LHX8* in HPV16 versus non-HPV16 controls of lavage self-samples (P value = 0.03; Supplementary Fig. S11).

ASCL1, achaete-scute family bHLH transcription factor 1, is a proneural transcription factor and functions as a main regulator of differentiation in neurogenesis[29]. *LHX8*, LIM homeobox 8, is a highly conserved transcription factor regulating cell fate in neurogenesis, tooth morphogenesis and oogenesis[30]. *ST6GALNAC5*, ST6 N-Acetylgalactosaminide Alpha-2,6-Sialyltransferase 5, is a transmembrane sialyltransferase involved in the biosynthesis of gangliosides on the cell surface[31]. Next to cervical cancer, *LHX8* methylation has been detected in breast cancer[32], *ST6GALNAC5* methylation has been described in colorectal cancer studies[33] and *ASCL1* methylation has been detected in oral and colorectal cancer[34,35].

Of the previously described DNA methylation markers tested in self-samples[12–14,17,24], the DNA methylation panel *FAM19A4/miR124-2* showed the best clinical performance in a large screening cohort. Analysis of the same study cohorts as used in the present study showed a CIN3+ sensitivity of 71% in lavage and 69% in brush self-samples at a specificity of 68% and 76%, respectively[14]. Within the CIN3+ group, 68% of CIN3 and all cancers were detected in both self-sample types. Other DNA methylation marker panels, such as *JAM3/EPB41L3/TERT/C13ORF18*, have only been analysed in small selected series of self-samples[15,16]. A combination of DNA methylation markers with HPV16/18 genotyping results in higher sensitivities compared with solely DNA methylation, however at the cost of severe lower specificities due to detection of early CIN2/3[14,36]. Our 3-gene methylation classifier shows a better sensitivity for CIN3 than other assays in both lavage (74%) and brush (88%) self-samples in a similar screening population, at a higher specificity of 79% and 81%, respectively. These findings emphasise the validity and importance of our approach to perform the DNA methylation marker discovery directly on self-sampled material. Furthermore, the 3-gene methylation classifier detected all self-samples from women with SCC. Importantly, all self-samples from women with SCC showed very high predicted probabilities (median, 1.00; range, 0.54-1.00), which accentuates the value

of our 3-gene methylation classifier for detection of cervical cancer. In addition, all self-samples from women with ACIS and AdCA scored DNA methylation-positive, indicating that glandular lesions are also detected by our 3-gene methylation classifier. Nevertheless, further evaluation of cervical glandular lesions and other rare cervical cancer types is warranted.

A limitation of our study is that we used cohorts of non-attending women. Therefore, further confirmation in a regular population-based population is warranted. In addition, the Infinium 450K array is limited to 485.577 CpG measurements. A new version of this platform, the Infinium MethylationEPIC Beadchip array, covers over 850,000 CpG sites and would yield more discovery data, especially in the enhancer regions[37]. Although the Infinium 450K array is not fully genome-wide and may yield partly different results than other methylome analysis methods, all 485,577 probes cover 99% of RefSeq genes and 96% of all CpG islands with multiple probes per gene and CpG island[38]. Furthermore, this array is one of the most widely accepted methods for genome-wide DNA methylation profiling and it is cost-effective[39].

In conclusion, by genome-wide DNA methylation profiling on self-samples obtained from a screening trial, we identified and validated an effective 3-gene methylation classifier for detection of CIN3 and cervical cancer in both lavage and brush self-samples from hrHPV-positive women. Moreover, this 3-gene methylation classifier showed an improved clinical performance compared with current (complex) triage strategies for the management of hrHPV-positive self-samples[13]. Our findings indicate that a transition toward full molecular self-screening in HPV-based cervical screening programmes is feasible.

ACKNOWLEDGEMENTS

We thank Lise De Strooper, Bart Hesselink, Maarten van der Salm, Saskia Doorn, Martijn Bogaarts and Dénira Agard for excellent technical assistance. In addition, we thank Dr. S. Farkas for providing the raw data of her study[22]. This work was supported by the European Research Council (ERC advanced 2012-AdG; 322986; Mass-Care) to C.J.L.M. Meijer and by ZonMw (Netherlands Organisation for Health Research and Development; 91216012) to M.A. van de Wiel.

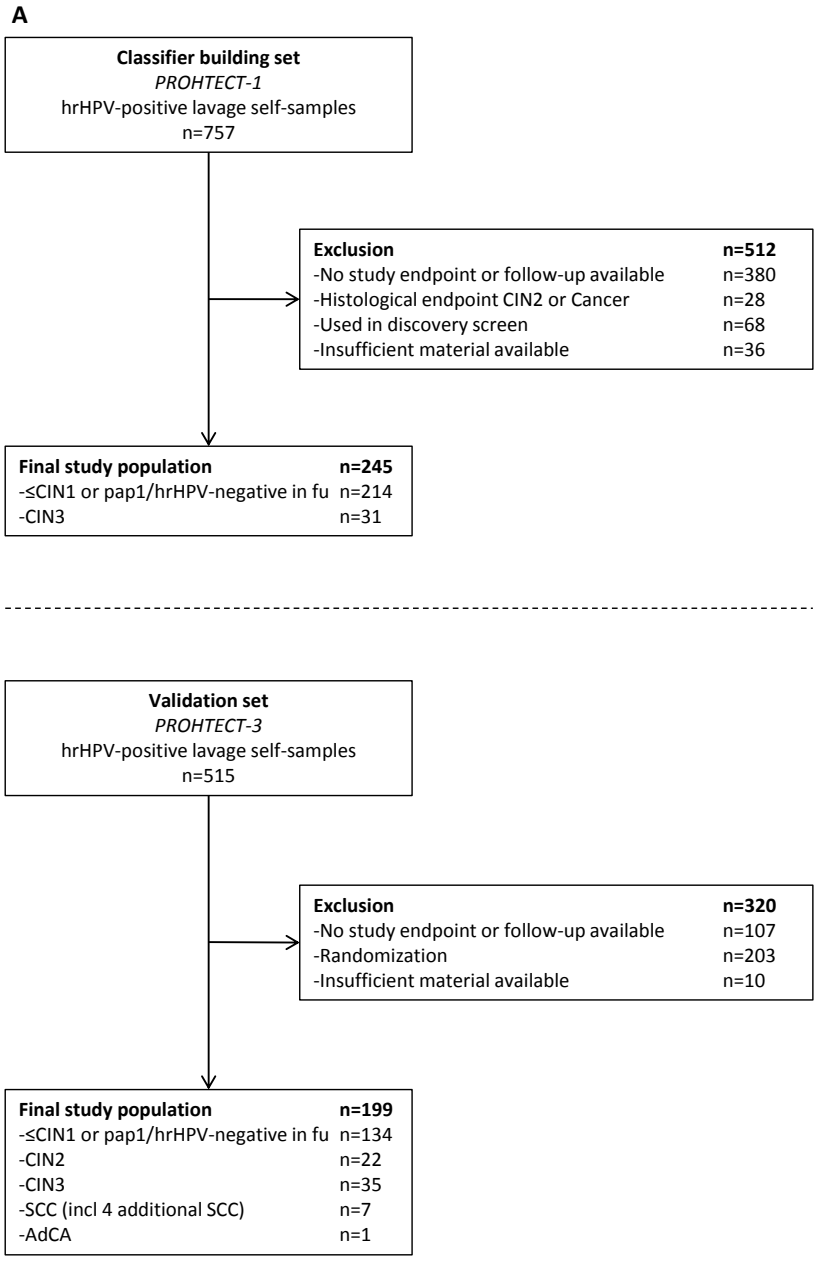
REFERENCES

1. Peto, J., Gilham, C., Fletcher, O. & Matthews, F. E. The cervical cancer epidemic that screening has prevented in the UK. *Lancet (London, England)* **364**, 249–56 (2004).
2. Bos, A. B., Rebolj, M., Habbema, J. D. F. & van Ballegooijen, M. Nonattendance is still the main limitation for the effectiveness of screening for cervical cancer in the Netherlands. *Int. J. cancer* **119**, 2372–5 (2006).
3. Gök, M. *et al.* HPV testing on self collected cervicovaginal lavage specimens as screening method for women who do not attend cervical screening: cohort study. *BMJ* **340**, c1040 (2010).
4. Gök, M. *et al.* Experience with high-risk human papillomavirus testing on vaginal brush-based self-samples of non-attendees of the cervical screening program. *Int. J. cancer* **130**, 1128–35 (2012).
5. Bais, A. G. *et al.* Human papillomavirus testing on self-sampled cervicovaginal brushes: An effective alternative to protect nonresponders in cervical screening programs. **120**, 1505–1510 (2007).
6. Racey, C. S., Withrow, D. R. & Gesink, D. Self-collected HPV Testing Improves Participation in Cervical Cancer Screening: A Systematic Review and Meta-analysis. 159–166 (2013).
7. Snijders, P. J. F. *et al.* High-risk HPV testing on self-sampled versus clinician-collected specimens: A review on the clinical accuracy and impact on population attendance in cervical cancer screening. *Int. J. Cancer* **132**, 2223–2236 (2013).
8. Arbyn, M. *et al.* Accuracy of human papillomavirus testing on self-collected versus clinician-collected samples: A meta-analysis. *Lancet Oncol.* **15**, 172–183 (2014).
9. Rijkaart, D. C. *et al.* Evaluation of 14 triage strategies for HPV DNA-positive women in population-based cervical screening. *Int. J. cancer* **130**, 602–10 (2012).
10. Dijkstra, M. G. *et al.* Primary hrHPV DNA testing in cervical cancer screening: how to manage screen-positive women? A POBASCAM trial substudy. *Cancer Epidemiol. Biomarkers Prev.* **23**, 55–63 (2014).
11. Garcia, F. *et al.* Cross-sectional Study of Patient- and Physician- Collected Cervical Cytology and Human Papillomavirus. **102**, 266–272 (2003).
12. Verhoef, V. M. J. *et al.* Triage by methylation-marker testing versus cytology in women who test HPV-positive on self-collected cervicovaginal specimens (PROHTECT-3): A randomised controlled non-inferiority trial. *Lancet Oncol.* **15**, 315–322 (2014).
13. Luttmer, R. *et al.* Management of high-risk HPV-positive women for detection of cervical (pre)cancer. *Expert Rev. Mol. Diagn.* **16**, 961–74 (2016).
14. De Strooper, L. M. A. *et al.* Validation of the FAM19A4/mir124-2 DNA methylation test for both lavage- and brush-based self-samples to detect cervical (pre)cancer in HPV-positive women. *Gynecol. Oncol.* **141**, 341–7 (2016).
15. Boers, A. *et al.* DNA methylation analysis in self-sampled brush material as a triage test in hrHPV-positive women. *Br. J. Cancer* **111**, 1095–101 (2014).
16. Eijsink, J. J. H. *et al.* Detection of cervical neoplasia by DNA methylation analysis in cervico-vaginal lavages, a feasibility study. *Gynecol. Oncol.* **120**, 280–283 (2011).

17. Hesselink, A. T. *et al.* Methylation marker analysis of self-sampled cervico-vaginal lavage specimens to triage high-risk HPV-positive women for colposcopy. *Int. J. cancer* **135**, 880–6 (2014).
18. Bosgraaf, R. P. *et al.* Comparative performance of novel self-sampling methods in detecting high-risk human papillomavirus in 30,130 women not attending cervical screening. *Int. J. cancer* **136**, 646–655 (2015).
19. Snellenberg, S. *et al.* Development of a multiplex methylation-specific PCR as candidate triage test for women with an HPV-positive cervical scrape. *BMC Cancer* **12**, 551 (2012).
20. Schmittgen, T. D. & Livak, K. J. Analysing real-time PCR data by the comparative CT method. *Nat. Protoc.* **3**, 1101–1108 (2008).
21. van de Wiel, M. A., Lien, T. G., Verlaat, W., van Wieringen, W. N. & Wilting, S. M. Better prediction by use of co-data: adaptive group-regularized ridge regression. *Stat. Med.* **35**, 368–381 (2016).
22. Farkas, S. A., Milutin-Gašperov, N., Grce, M. & Nilsson, T. K. Genome-wide DNA methylation assay reveals novel candidate biomarker genes in cervical cancer. *Epigenetics* **8**, 1213–1225 (2013).
23. Christensen, B. C. *et al.* Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CPG island context. *PLoS Genet.* **5**, e1000602 (2009).
24. Steenbergen, R. D. M., Snijders, P. J. F., Heideman, D. A. M. & Meijer, C. J. L. M. Clinical implications of (epi)genetic changes in HPV-induced cervical precancerous lesions. *Nat. Rev. Cancer* **14**, 395–405 (2014).
25. De Strooper, L. M. A. *et al.* Methylation analysis of the FAM19A4 gene in cervical scrapes is highly efficient in detecting cervical carcinomas and advanced CIN2/3 lesions. *Cancer Prev. Res.* **7**, 1251–1257 (2014).
26. Novianti, P. W., Snoek, B. C., Wilting, S. M. & van de Wiel, M. A. Better diagnostic signatures from RNAseq data through use of auxiliary co-data. *Bioinformatics* **33**, btw837 (2017).
27. Boers, A. *et al.* Discovery of new methylation markers to improve screening for cervical intraepithelial neoplasia grade 2/3. *Clin. Epigenetics* **8**, 29 (2016).
28. Clarke, M. A. *et al.* Discovery and validation of candidate host DNA methylation markers for detection of cervical precancer and cancer. *Int. J. Cancer* **141**, 701–710 (2017).
29. Vasconcelos, F. F. & Castro, D. S. Transcriptional control of vertebrate neurogenesis by the proneural factor Ascl1. *Front. Cell. Neurosci.* **8**, 412 (2014).
30. Zhou, C. *et al.* Lhx6 and Lhx8: Cell fate regulators and beyond. *FASEB J.* **29**, 4083–4091 (2015).
31. Drolez, A. *et al.* ST6GALNAC5 expression decreases the interactions between breast cancer cells and the human blood-brain barrier. *Int. J. Mol. Sci.* **17**, 1309 (2016).
32. Tommasi, S., Karm, D. L., Wu, X., Yen, Y. & Pfeifer, G. P. Methylation of homeobox genes is a frequent and early epigenetic event in breast cancer. *Breast cancer Res.* **11**, R14 (2009).
33. Øster, B. *et al.* Identification and validation of highly frequent CpG island hypermethylation in colorectal adenomas and carcinomas. *Int. J. Cancer* **129**, 2855–2866 (2011).
34. Jin, B. *et al.* DNMT1 and DNMT3B modulate distinct polycomb-mediated histone modifications in colon cancer. *Cancer Res.* **69**, 7412–7421 (2009).

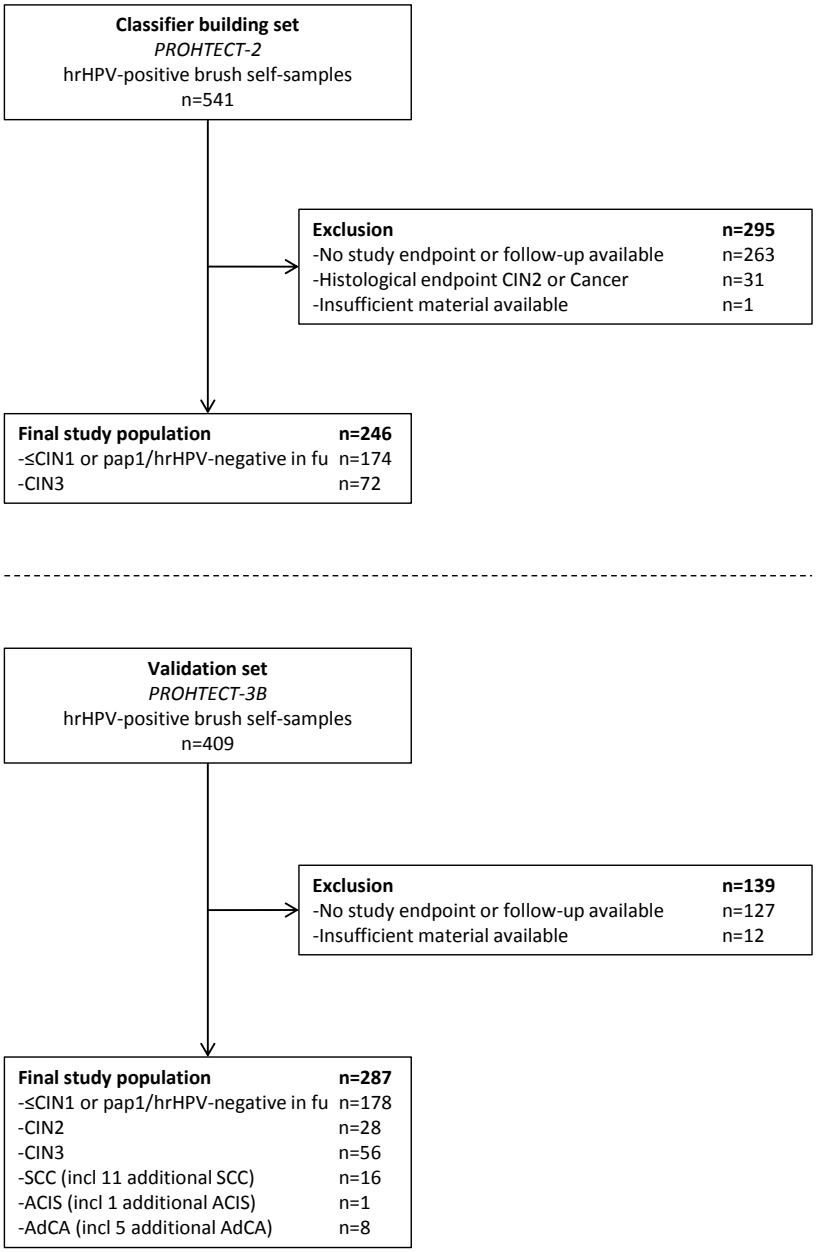
35. Li, Y.-F. *et al.* DNA methylation profiles and biomarkers of oral squamous cell carcinoma. *Epigenetics* **10**, 229–236 (2015).
36. Verhoef, V. M. J. *et al.* Methylation marker analysis and HPV16/18 genotyping in high-risk HPV positive self-sampled specimens to identify women with high grade CIN or cervical cancer. *Gynecol. Oncol.* **135**, 58–63 (2014).
37. Moran, S., Arribas, C. & Esteller, M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* **8**, 389–399 (2016).
38. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295 (2011).
39. Yong, W.-S., Hsu, F.-M. & Chen, P.-Y. Profiling genome-wide DNA methylation. *Epigenetics Chromatin* **9**, 26 (2016).

SUPPLEMENTARY FIGURES AND TABLES

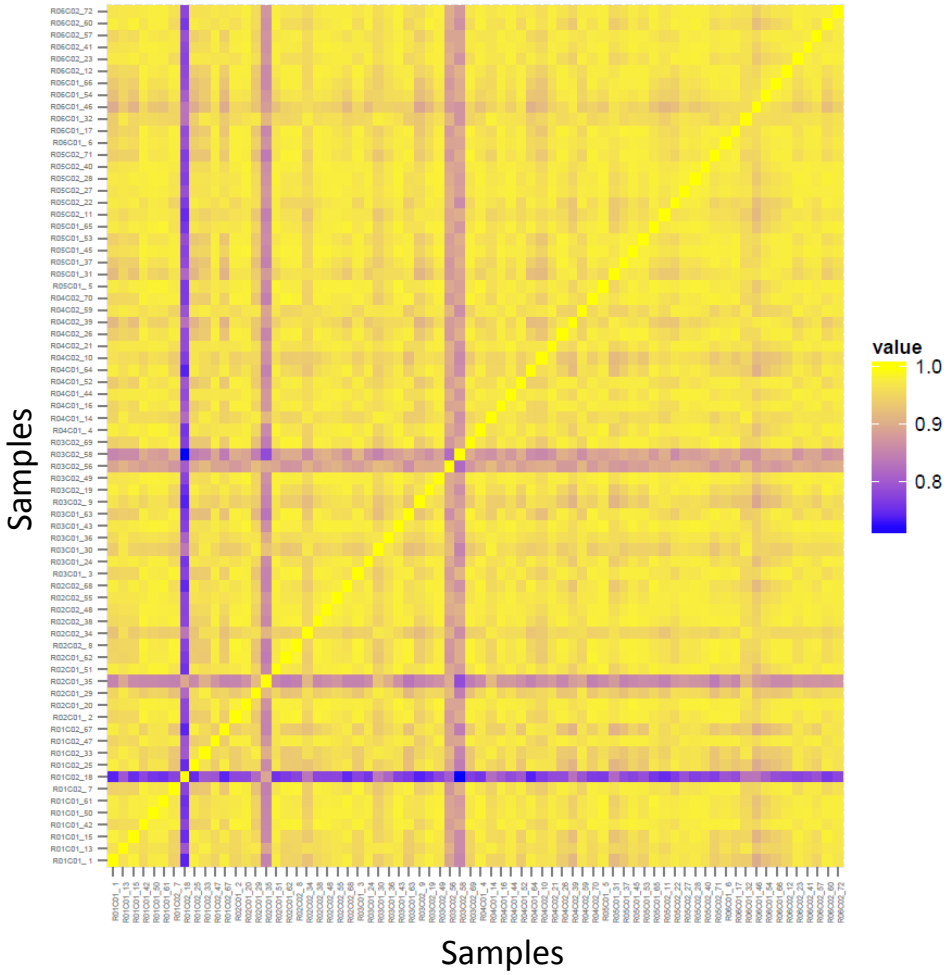


Supplementary Figure S1. Overview of study populations. The classifier building sets and validation sets for (**A**) hrHPV-positive lavage self-samples and (**B**) hrHPV-positive brush self-samples. fu: follow-up.

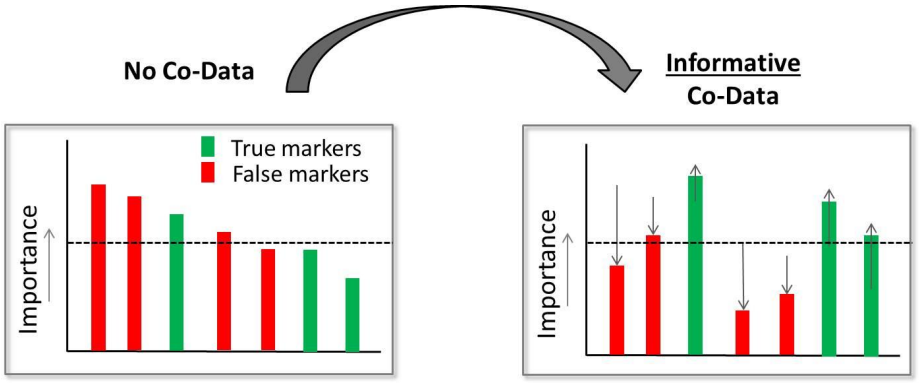
B



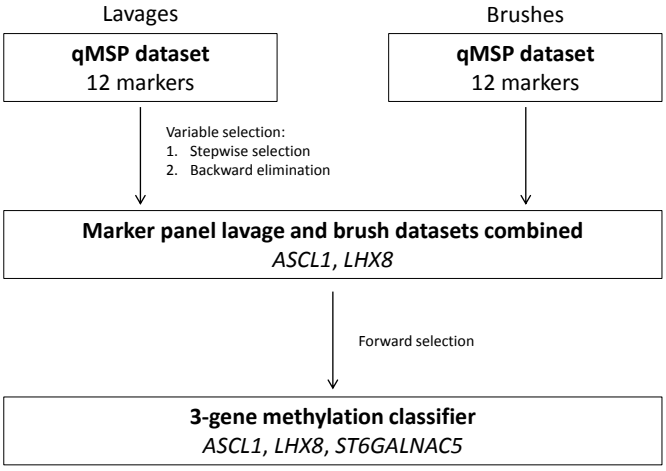
Supplementary Figure S1. Continued.



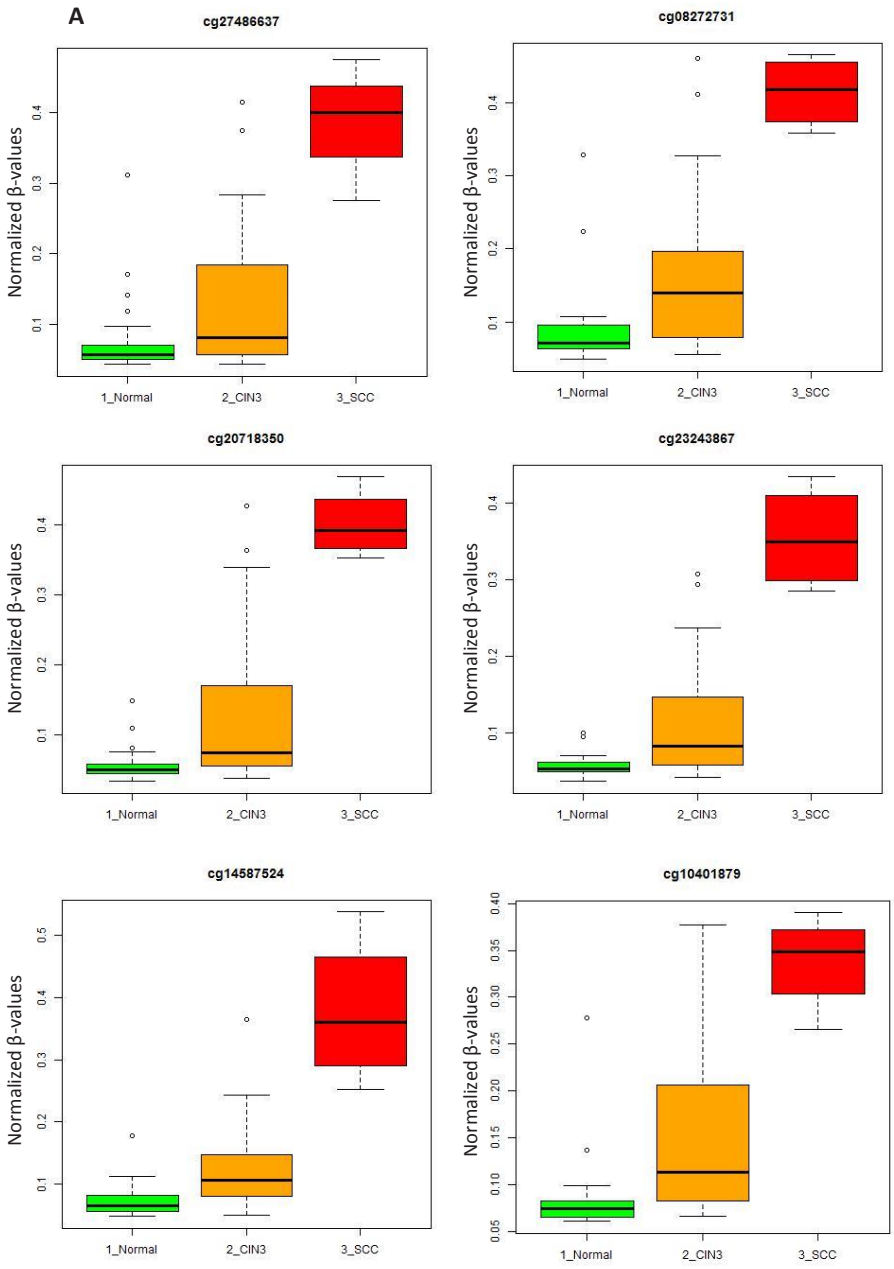
Supplementary Figure S2. Correlation heatmap of all 72 hrHPV-positive self-samples analysed by Infinium Methylation 450K Array. Both x-axis and y-axis represent all 72 samples, which are compared with each other showing high (yellow) or low (blue) correlation.



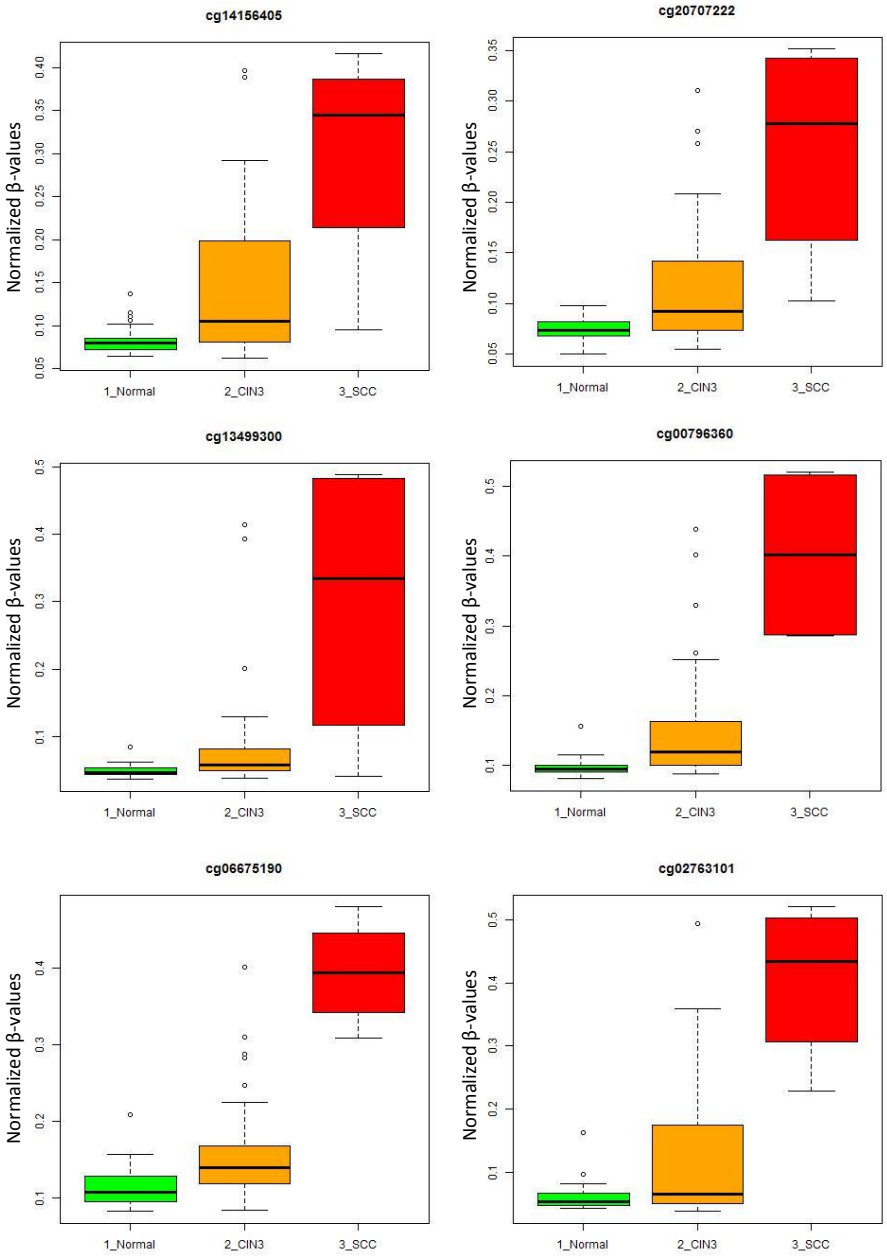
Supplementary Figure S3. The effect of informative co-data. Using informative co-data, the weights help enhancing the variables truly related to the outcome while suppressing the false variables.



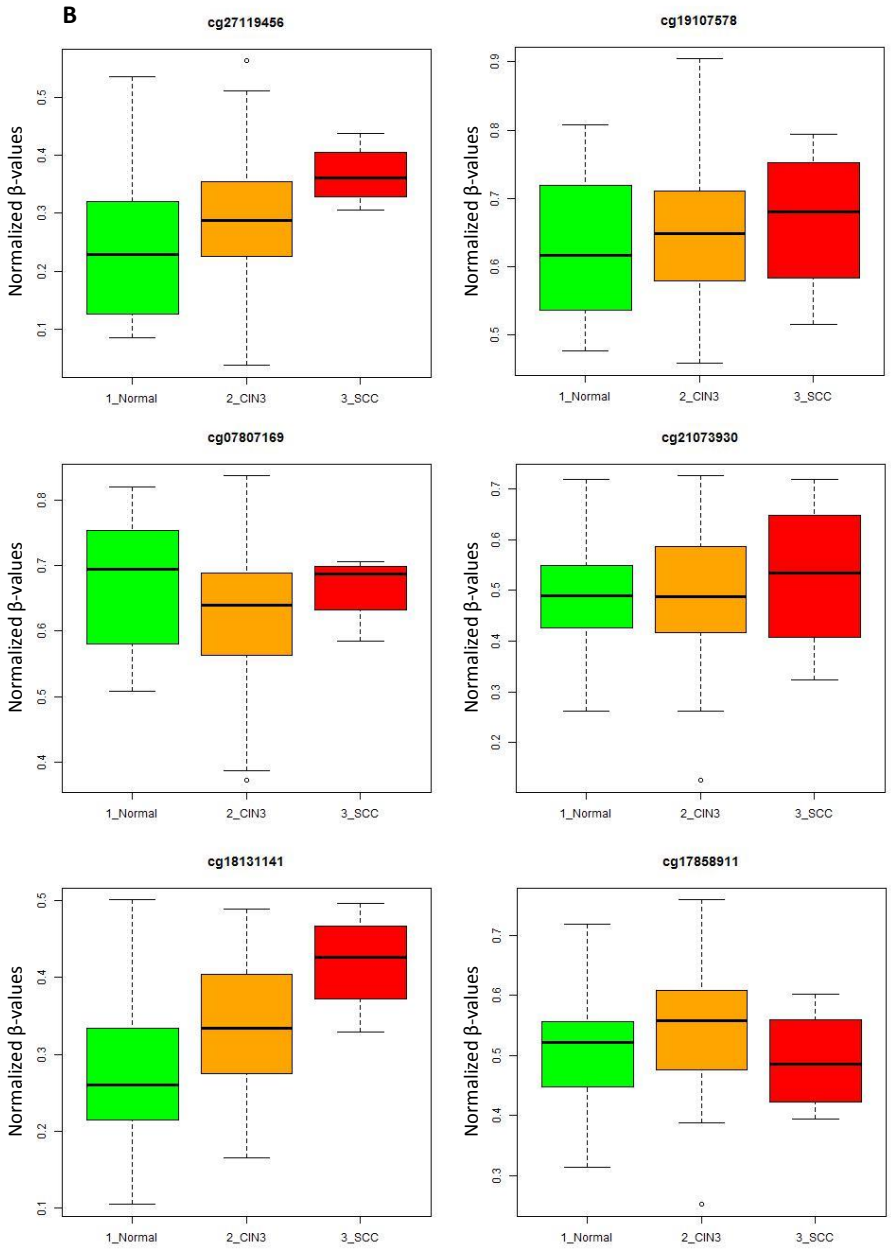
Supplementary Figure S4. Approach to build a 3-gene methylation classifier applicable to both lavage and brush self-samples.



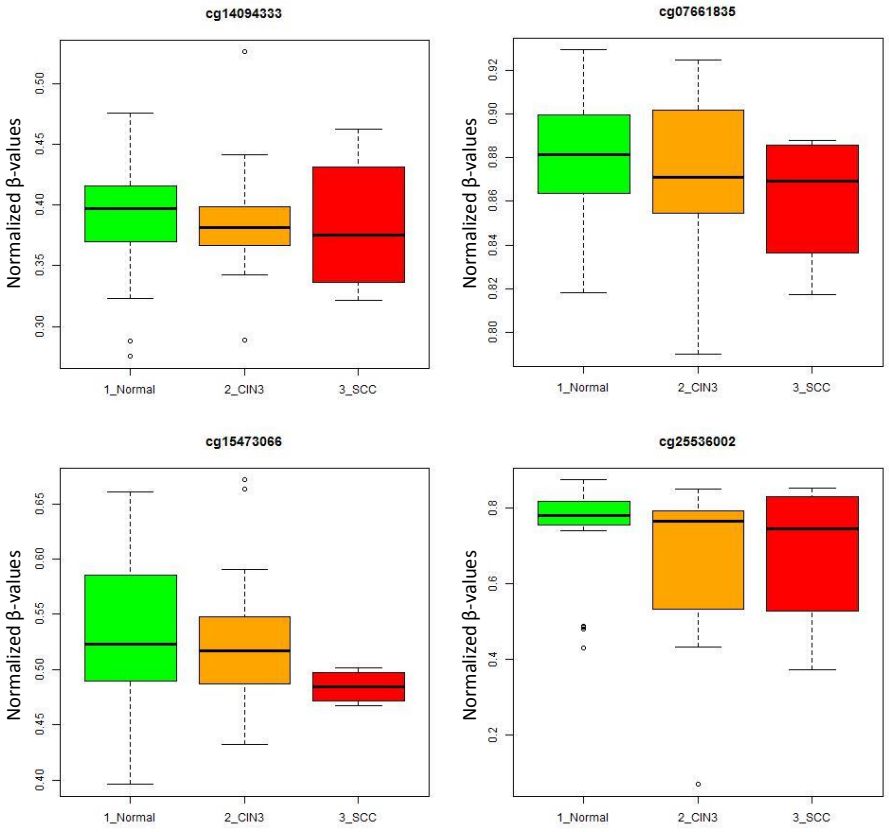
Supplementary Figure S5. DNA methylation levels of all 28 DNA methylation markers in hrHPV-positive lavage self-samples. Normalised β-values (y-axis) from the Infinium Methylation 450K array in hrHPV-positive controls (Normal; green; n=28), CIN3 (orange; n=36) and SCC (red; n=4) lavage self-samples for (**A**) the 12 selected most discriminative DNA methylation markers and (**B**) the 16 remaining DNA methylation markers.



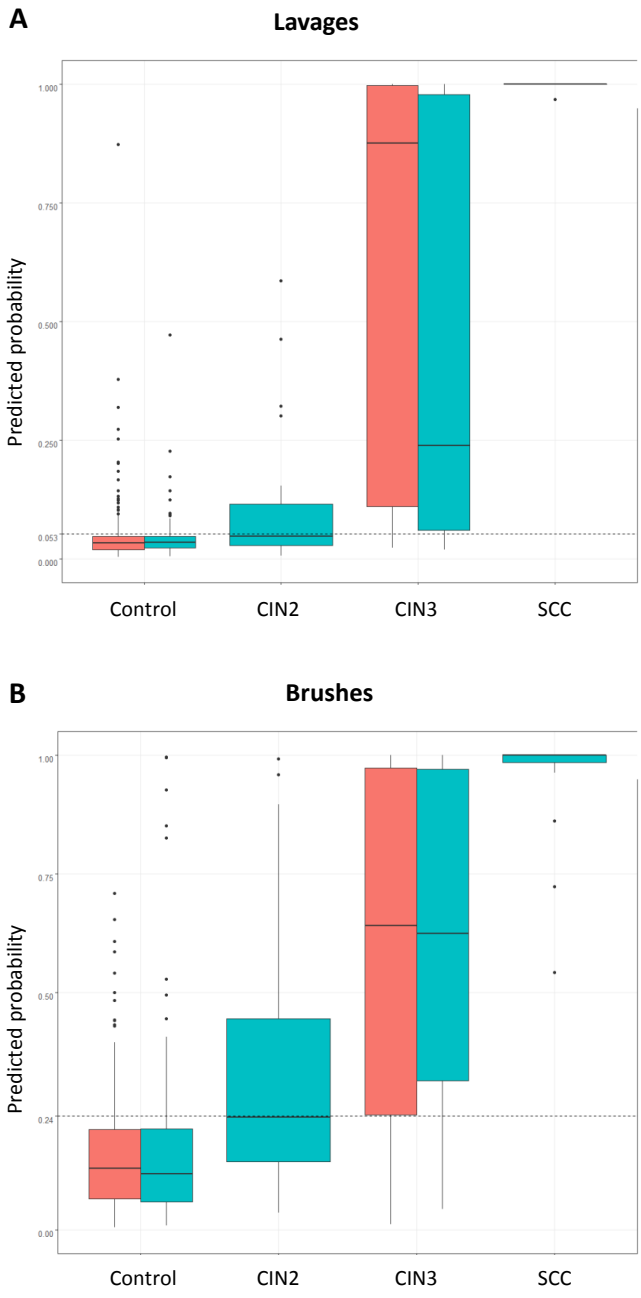
Supplementary Figure S5. Continued.



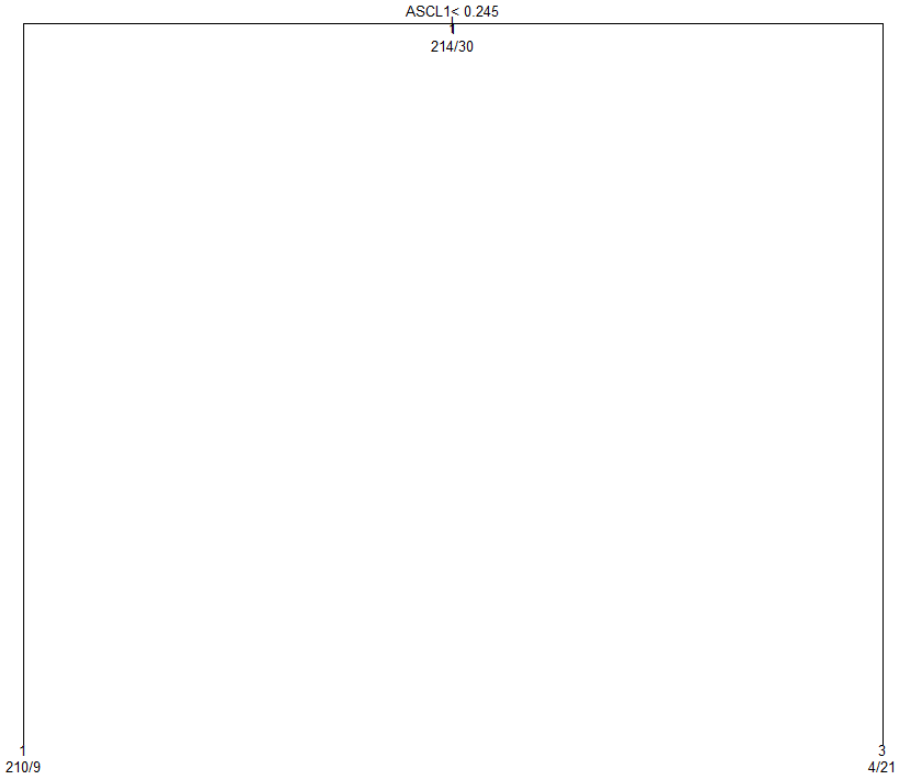
Supplementary Figure S5. Continued.



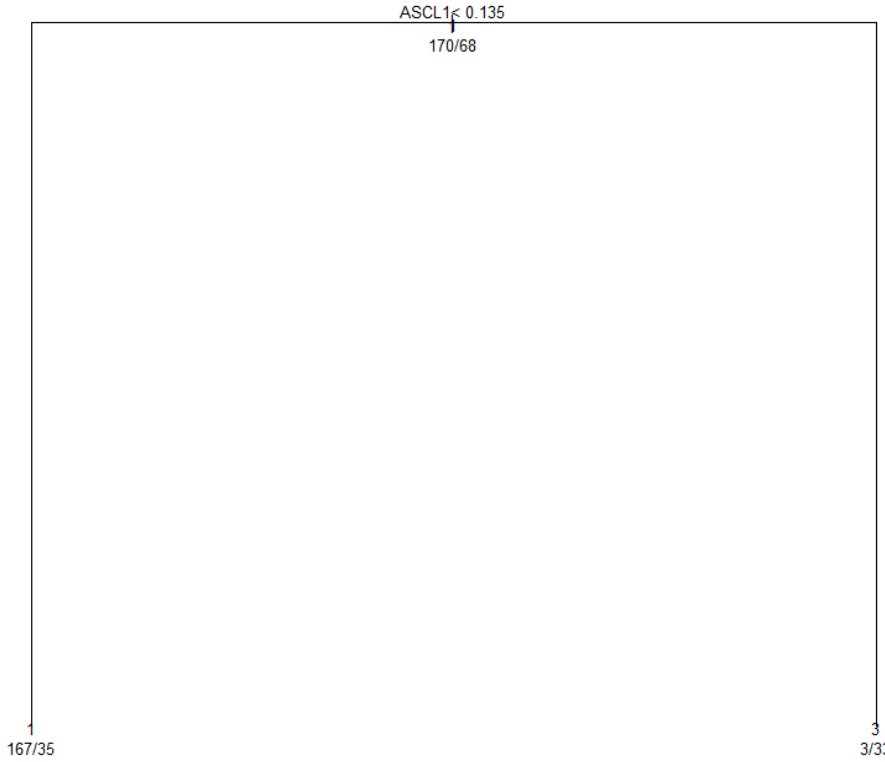
Supplementary Figure S5. Continued.



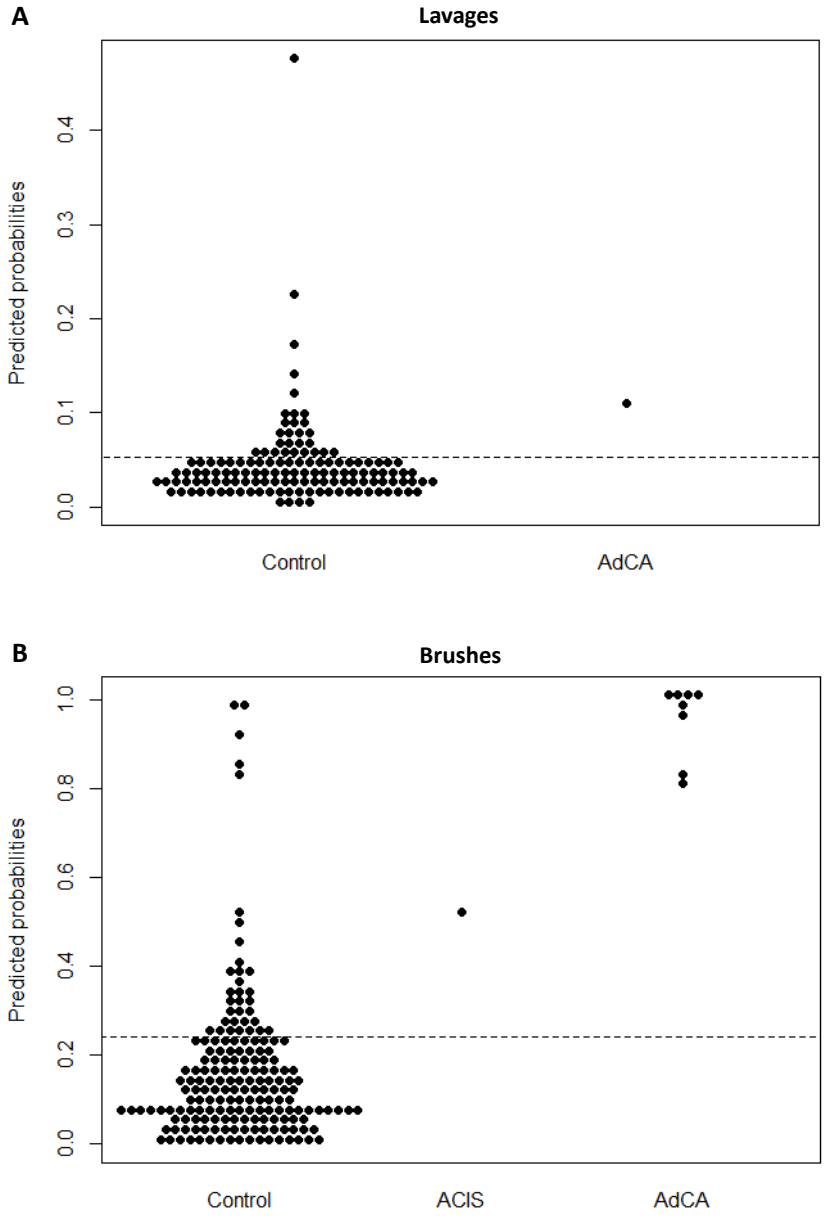
Supplementary Figure S6. DNA methylation-positivity determined by the 3-gene methylation classifier. Predicted probabilities of all tested hrHPV-positive (**A**) lavage and (**B**) brush self-samples in the classifier building set (red) and validation set (blue). The dashed black line indicates the cut-off used for DNA methylation-positivity.



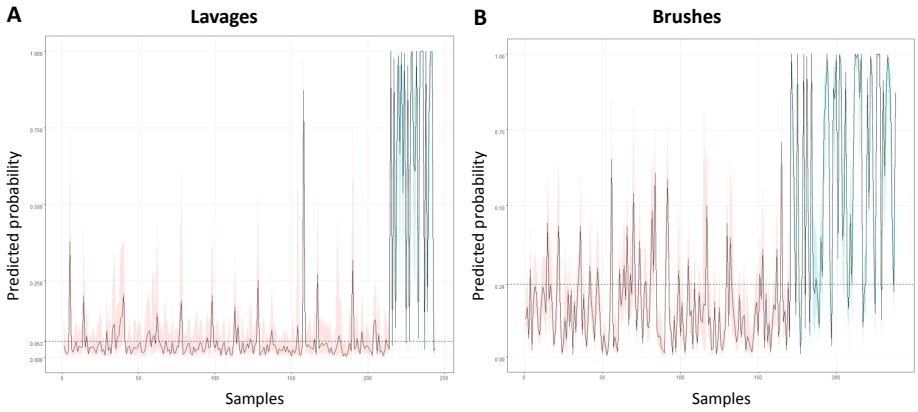
Supplementary Figure S7. Classification tree model for lavage self-samples. The CART analysis in lavage self-samples resulted in a tree model consisting of only *ASCL1*. Using a cut-off of 0.245, 210 of 214 hrHPV-positive controls were correctly classified as controls (1) and 21 of 30 CIN3 were correctly classified as CIN3 (3).



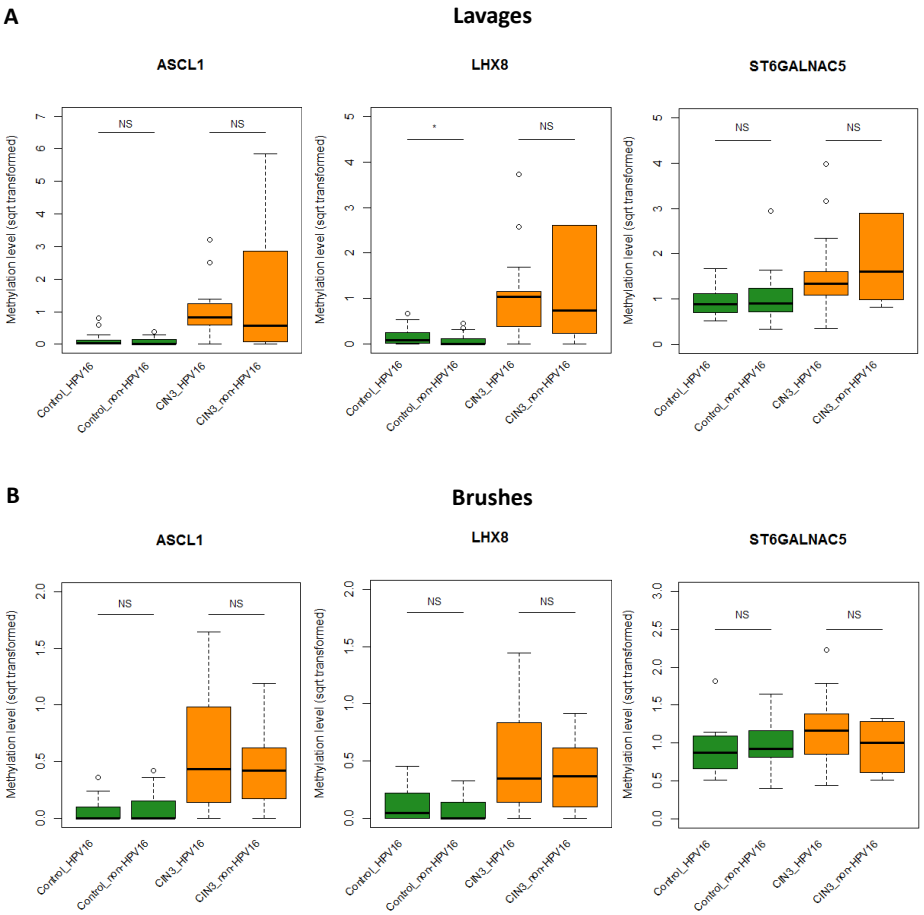
Supplementary Figure S8. Classification tree model for brush self-samples. The CART analysis in brush self-samples resulted in a tree model consisting of only *ASCL1*. Using a cut-off of 0.135, 167 of 170 hrHPV-positive controls were correctly classified as controls (1) and 33 of 68 CIN3 were correctly classified as CIN3 (3).



Supplementary Figure S9. DNA methylation-positivity in hrHPV-positive self-samples from women with glandular lesions and cancer. (**A**) Predicted probabilities in lavage self-samples from hrHPV-positive controls in the validation set (n=134) and from a woman with an adenocarcinoma (AdCA; n=1). (**B**) Predicted probabilities in brush self-samples from hrHPV-positive controls in the validation set (n=178) and from women with adenocarcinoma in situ (ACIS; n=1) or AdCA (n=8). The dashed line indicates the cut-off used for DNA methylation-positivity.



Supplementary Figure S10. Predicted probabilities in self-samples from women with and without CIN3. Predicted probabilities are shown for hrHPV-positive (**A**) lavage and (**B**) brush self-samples in the classifier building set. The 95% confidence interval is shown as a coloured range around the predicted probability for samples from hrHPV-positive controls (pink) and women with CIN3 (blue). The dashed black line indicates the cut-off used for DNA methylation-positivity.



Supplementary Figure S11. DNA methylation levels of *ASCL1*, *LHX8* and *ST6GALNAC5* in self-samples from women with HPV16 or with other hrHPV types (non-HPV16). The DNA methylation levels (square root-transformed) are shown for (**A**) lavage and (**B**) brush self-samples in a subset of the classifier building set, which had available data of hrHPV types. The hrHPV-positive controls and CIN3 were divided in self-samples with HPV16 and non-HPV16. The lavage self-sample series comprised n=16 controls with HPV16, n=64 controls with non-HPV16, n=13 CIN3 with HPV16 and n=6 CIN3 with non-HPV16. The brush self-sample series comprised n=10 controls with HPV16, n=52 controls with non-HPV16, n=24 CIN3 with HPV16 and n=10 CIN3 with non-HPV16. * $P < 0.05$; NS: not significant.

Supplementary Table S1. Primers and probes used in multiplex qMSP.

DNA methylation marker	bp length	Anealing temp (°C)	Amplicon size (bp)
ACAN-F	30	59.3	
ACAN-R	24	60.8	117
ACAN-P	35	69.7	
ASCL1-F	28	60.1	
ASCL1-R	24	60.9	117
ASCL1-P	36	69.9	
LHX8-F	21	59.6	
LHX8-R	23	58.2	110
LHX8-P	38	68.3	
MYADM-F	20	58.8	
MYADM-R	22	56.8	112
MYADM-P	36	68.7	
NRG3-F	20	60.0	
NRG3-R	23	59.1	101
NRG3-P	25	68.4	
RGS7-F	22	59.4	
RGS7-R	24	59.6	92
RGS7-P	34	68.3	
ST6GALNAC3-F	24	58.9	
ST6GALNAC3-R	23	61.5	124
ST6GALNAC3-P	31	69.8	
ST6GALNAC5-F	26	60.1	
ST6GALNAC5-R	23	61.4	104
ST6GALNAC5-P	20	69.6	
WDR17-F	17	59.1	
WDR17-R	25	60.9	82
WDR17-P	27	69.8	
ZNF582-F	25	58.5	
ZNF582-R	23	58.2	99
ZNF582-P	30	69.4	
ZNF583-F	31	60.0	
ZNF583-R	17	58.5	105
ZNF583-P	29	68.6	
ZNF781-F	20	60.9	
ZNF781-R	24	58.9	76
ZNF781-P	25	69.5	
ACTB-F	25	58.2	
ACTB-R	27	58.9	133
ACTB-P	30	68.9	

Sequences are available on request. F: forward primer; R: reverse primer; P: probe.

Supplementary Table S2. Logistic regression analysis results of the 3-gene methylation classifier for CIN3 detection in hrHPV-positive self-samples.

	Lavage self-samples		
3-gene methylation classifier	**P value[a]**	**Coefficient/sd[b]**	**AUC**
ASCL1	0.028	2.050	
LHX8	0.009	2.332	0.90
ST6GALNAC5	0.572	-0.549	
age	0.041	-1.880	
Without age			
ASCL1	0.006	2.576	
LHX8	0.019	2.186	0.88
ST6GALNAC5	0.208	-1.200	
Without *ST6GALNAC5*			
ASCL1	0.034	2.011	
LHX8	0.009	2.336	0.89
age	0.020	-2.140	

[a] P value of the corresponding DNA methylation marker in the logistic regression model.

[b] Coefficient divided by the standard deviation (sd) of each DNA methylation marker in the logistic regression model.

Supplementary Table S3. The comparison of the performance of logistic regression and CART models.

Model	Data	Self-sample type	Accuracy
CART	Training	Lavages	94.67%
		Brushes	84.03%
	Validation	Lavages	84.03%
		Brushes	85.04%
Logistic regression	Training	Lavages	94.67%
		Brushes	84.87%
	Validation	Lavages	88.76%
		Brushes	85.89%

Accuracy (i.e. the proportion of the number of correctly classified samples with the total number of samples), sensitivity, negative predictive value (NPV) and positive predictive value (PPV) on logistic regression models were based on a cut-off for the corresponding specificity of CART model on the training set, which was 0.262 and 0.630 for lavage and brush self-samples, respectively.

Brush self-samples		
P value[a]	Coefficient/sd[b]	AUC
< 0.001	4.149	
0.012	2.638	0.86
0.041	-1.989	
< 0.001	-3.778	
< 0.001	4.296	
0.020	2.412	0.81
0.006	-2.629	
< 0.001	3.926	
0.023	2.397	0.84
< 0.001	-4.061	

Sensitivity	Specificity	NPV	PPV
70.00%	98.13%	95.89%	84.00%
48.53%	98.24%	82.67%	91.67%
45.71%	99.25%	87.50%	94.12%
46.43%	97.19%	85.22%	83.87%
70.00%	98.13%	98.13%	70.00%
50.00%	98.82%	98.82%	50.00%
48.57%	99.25%	88.08%	94.44%
50.00%	97.19%	86.07%	84.85%

SUPPLEMENTARY METHODS

Infinium HumanMethylation450 BeadChip data preprocessing

Probes which, (i) showed a low bead count (< 3 in at least 5% of samples), (ii) had a detection P value of > 0.05 in at least 3% of samples, (iii) contained SNPs at or within 10 bp from the target CpG-site[1], (iv) were vulnerable to cross-hybridisation[1] or (v) were located on allosomes, were removed, leaving 365,620 probes for analysis. Data from the remaining 68 samples were normalised using dasen normalisation[2], which uses a different background correction and quantile normalisation for type I and type II probes. It returns normalised beta-values, representing methylation percentage at the corresponding CpG-site. These beta-values were transformed to a Gaussian scale by the commonly used arcsine square root transformation. As an extra filter, we selected probes that (i) were related to a gene according to the Illumina annotation file (v1.2), (ii) were related to a CpG-island, -shore or -shelf according to the Illumina annotation file (v1.2) and (iii) showed increased methylation in CIN3 compared with hrHPV-positive controls (difference in β-value ≥ 0.1) in an independent tissue-based study[3].

DNA isolation, hrHPV testing and bisulphite treatment

DNA from self-samples was isolated using the NucleoMag 96 Tissue kit (Macherey-Nagel, Düren, Germany) and a Microlab Star robotic system (Hamilton, Martinsried, Germany) according to the manufacturer's protocol. DNA concentration was measured using the NanoDrop ND-1000 (Thermo Scientific, Waltham, MA, USA). In PROHTECT-1 and PROHTECT-2 trials, hrHPV detection was performed by Hybrid Capture-2 (HC2; Qiagen, Hilden, Germany) as described previously[4,5]. HrHPV types were determined in a subset of PROHTECT-1 and PROHTECT-2 samples using the GP5+/6+ RLB system[6]. Samples from the PROHTECT-3 and PROHTECT-3B trials were tested for hrHPV by GP5+/6+ PCR using the Diassay EIA HPV GP HR kit (Diassay, Voorburg, The Netherlands) as described previously[7,8]. HrHPV detection in samples from the X15MET study was performed by the HPV-Risk assay (Self-screen B.V., Amsterdam, The Netherlands)[9]. Prior to DNA methylation analysis, isolated DNA was bisulphite-converted using the EZ DNA Methylation kit (Zymo Research, Orange, CA, USA).

DNA methylation analysis by multiplex qMSP

For multiplex qMSP, EpiTect MethyLite Master Mix (Qiagen), 50ng of bisulphite-converted DNA and 100-300nM of each primer and fluorescent dye-labelled probe was used. Primer and probe information is shown in Supplementary Table S1. All multiplex qMSP analyses were performed on the ViiA7 (Applied Biosystems, Foster City, CA, USA). A plasmid containing the amplicon sequences of all targets and *ACTB*, was used as calibrator. This calibrator functions as a technical quality control and was used to calculate $\Delta\Delta C_t$ ratios. Cycle threshold (C_t) values were measured at a fixed fluorescence threshold. All lab work was performed blinded to the results.

High-dimensional data analysis on DNA methylation array data

We applied adaptive group-regularised (logistic) ridge regression[10,11]. GRridge can be regarded as a weighted regression for omics data. As in commonly used ridge or lasso models, regression coefficients are down-weighted ('penalised') to account for the high-dimensionality of the data. However, our approach allows the use of prior information ('co-data') on the genomic variables to differentiate the weights, which can come from internal and external sources. Unlike meta-analysis, the co-data complements the primary study data; the latter will still be the main driver behind the classifier. Moreover, the co-data are only required during training of the classifier, not when applying the classifier to new (test) samples.

The weights are estimated in an objective, unbiased way using so-called empirical Bayes techniques[10]. Roughly, this measures how relevant the co-data is for the primary data at hand, and if so, allows more skewed weights across groups of variables. If the co-data is not relevant, it will automatically assign flat weights, which means that the performance then converges to the classical logistic ridge regression model. In our study, we incorporated co-data from external and internal sources. For the first co-data, we used information from a similar DNA methylation study in cervical cancer[3]. For each and every probe in this external set, we computed a *P* value from the moderated t-statistic[12]. Next, we grouped the probes in our primary dataset based on these *P* values into a partition of 100 groups and forced monotony for the group-specific penalties, meaning that a group of variables with higher *P* values than another group will receive a penalty that is at least as high as that of the other group. This highly stabilises the weights. If the partition is informative to the primary data set, the probes with a low *P* value (high DNA methylation signal) would receive more weight. As second source of co-data, we used the standard deviation (sd) of the probes in our primary dataset. The sd was calculated across samples, ignoring the class labels (or responses). The use of sd as co-data was motivated in[10].

Having estimated the group-penalties, a weighted logistic regression is used as a classifier. We showed that the weights aid in improving predictive performance in particular when a parsimonious model (with few genomics variables) is desired, as is often the case in clinical settings. The philosophy is that the weights help enhancing the variables truly related to the outcome while suppressing the false variables (Supplementary Fig. S3). It is important to mention that the co-data should be defined independently from the response in the training set, in order to avoid overfitting.

Logistic regression analysis to build a DNA methylation classifier

Here, we employed the classical logistic regression method to solve the binary classification case with the previously selected methylation targets from the GRridge model.

On each brush ($\mathcal{B}$)- and lavage ($\mathcal{L}$) self-samples dataset, we built a (classical) logistic regression model. To compensate for increased DNA methylation due to age, the regression models were corrected by Age. For an independent response from a sample j on the $\mathcal{B}$ dataset (denoted as $Y_j^{\mathcal{B}}$, $j = 1, \dots, n_{\mathcal{B}}$; $n_{\mathcal{B}}$ is the number of samples on brush self-samples dataset), the model by incorporating k $(k = 1, \dots, 12)$ methylation markers can be written as follow:

$$Y_j^{\mathcal{B}} = Bernoulli\left(p_j^{\mathcal{B}}\right);\ logit\left(p_j^{\mathcal{B}}\right) = \eta_j^{\mathcal{B}}$$

$$\eta_j^{\mathcal{B}} = \beta_0^{\mathcal{B}} + \sum_{i=1}^{k} \beta_i^{\mathcal{B}} x_{ij}^{\mathcal{B}} + \beta_{age}^{\mathcal{B}} Age_j \quad (1)$$

Similarly, the model for lavage-self dataset is written as:

$$\eta_j^{\mathcal{L}} = \beta_0^{\mathcal{L}} + \sum_{i=1}^{k} \beta_i^{\mathcal{L}} x_{ij}^{\mathcal{L}} + \beta_{age}^{\mathcal{L}} Age_j \quad (2)$$

where p is the probability of a woman to be diagnosed with a CIN3, and; x_{ij} is the $\Delta\Delta C_t$ ratios of methylation marker i from sample j (transformed to the square root scale).

We checked interaction effects in the logistic regression model of each dataset. First, omnibus test for the interaction effects was conducted. A likelihood-ratio test was applied to test the difference of fits from two nested models, i.e. a model with all possible interaction terms and a model with main effects only. This test evaluates whether adding interaction terms significantly improves the performance of the model.

As both datasets agreed that interaction terms did not significantly improve the goodness-of-fit of logistic regression models (P value=0.136 and 0.199 for brush and lavage self-samples dataset, respectively), we further processed the predictive modeling by neglecting interaction terms on the model.

We built logistic regression models such that the selected markers improved the performance of regression models in *both* datasets. As a starting point, we incorporated all markers ($k = 12$) to model (1) and model (2). We then estimated joined-AIC (Akaike information criterion) from both models by:

$$-2\frac{n_{\mathcal{L}}}{n_{\mathcal{L}}+n_{\mathcal{B}}}logLik_{(1)}-2\frac{n_{\mathcal{B}}}{n_{\mathcal{L}}+n_{\mathcal{B}}}logLik_{(2)}+2df$$

where $logLik_{(1)}$ and $logLik_{(2)}$ are the log-likelihoods from model (1) and model (2), respectively, and; df (degree of freedom) is the number of estimated parameters. Backward elimination was applied to exclude less significant markers, by comparing joined-AICs between two nested models, i.e. the regression model with k and $k-1$ markers.

In addition, a standard AIC-based stepwise selection algorithm (as implemented in the 'stepAIC' function of R software) was applied to the regression models to allow additional relevant markers for CIN3 detection in either of the two separate datasets, as motivated in the main article.

Classification and regression tree

We applied an algorithm-based method, named classification and regression tree (CART) method, to do a binary classification on healthy control (hrHPV-positive controls) and case (CIN3) samples in the classifier building set. A tree model was grown by applying gini index as a measurement of node impurity. Further, once a complete model was built, the tree was pruned by optimising a cost parameter from adding a variable (denoted by α) with cross-validation procedure. More formally, the cost for the tree is written as follows[13],

$$R_{\alpha}(T)=R(T)+\alpha|T|$$

where $|T|$ is the number of terminal nodes and $R(T)$ is risk of a tree T, $R(T)=\sum_{i=1}^{k}P(T_i)R(T_i)$, with $T_1, T_2, \ldots, T_k$ are the terminal nodes of a tree T and $P(T_i)$ is probability of terminal node i to be included in the tree. To note, $R(T)$ can be seen as the residual sum of squares in a regression model. We refer to Breiman et al.[13] for further details of CART method. We used rpart package for this tree predictive modeling[14]. The CART analysis resulted in a tree consisting of *ASCL1* for both lavage and brush self-samples (Supplementary Fig. S7 and S8).

The results were compared by those of logistic regression by assessing accuracy (i.e. the proportion of the number of correctly classified samples with the total number of samples) and sensitivity at given specificity, the latter being fixed by the CART result.

REFERENCES

1. Chen, Y. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).
2. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. **14**, (2013).
3. Farkas, S. A., Milutin-Gašperov, N., Grce, M. & Nilsson, T. K. Genome-wide DNA methylation assay reveals novel candidate biomarker genes in cervical cancer. *Epigenetics* **8**, 1213–1225 (2013).
4. Gök, M. *et al.* HPV testing on self collected cervicovaginal lavage specimens as screening method for women who do not attend cervical screening: cohort study. *BMJ* **340**, c1040 (2010).
5. Gök, M. *et al.* Experience with high-risk human papillomavirus testing on vaginal brush-based self-samples of non-attendees of the cervical screening program. *Int. J. cancer* **130**, 1128–35 (2012).
6. Brule, A. J. C. Van Den *et al.* GP5+/6+ PCR followed by Reverse Line Blot Analysis Enables Rapid and High-Throughput Identification of Human Papillomavirus Genotypes. *J. Clin. Microbiol.* **40**, 779–787 (2002).
7. Verhoef, V. M. J. *et al.* Triage by methylation-marker testing versus cytology in women who test HPV-positive on self-collected cervicovaginal specimens (PROHTECT-3): A randomised controlled non-inferiority trial. *Lancet Oncol.* **15**, 315–322 (2014).
8. Bosgraaf, R. P. *et al.* Comparative performance of novel self-sampling methods in detecting high-risk human papillomavirus in 30,130 women not attending cervical screening. *Int. J. cancer* **136**, 646–655 (2015).
9. Hesselink, A. T. *et al.* Clinical Validation of the HPV-Risk Assay, a Novel Real-Time PCR Assay for Detection of High-Risk Human Papillomavirus DNA by Targeting the E7 Region. *J. Clin. Microbiol.* **52**, 890–896 (2014).
10. van de Wiel, M. A., Lien, T. G., Verlaat, W., van Wieringen, W. N. & Wilting, S. M. Better prediction by use of co-data: adaptive group-regularized ridge regression. *Stat. Med.* **35**, 368–381 (2016).
11. Novianti, P. W., Snoek, B. C., Wilting, S. M. & van de Wiel, M. A. Better diagnostic signatures from RNAseq data through use of auxiliary co-data. *Bioinformatics* **33**, btw837 (2017).
12. Smyth, G. K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Art. 3 (2004).
13. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and Regression Trees. The Wadsworth Statistics/Probability series* (1984).
14. Terry Therneau, Beth Atkinson and Brian Ripley. rpart: Recursive Partitioning and Regression Trees. https://CRAN.R-project.org/package=rpart. Accessed 2017.